У роботі запропонована модель семантичного контексту текстових масивів. Показано, що на основі решітки семантичних концептів можна сформувати семантичний базис для групування текстів за допомогою ієрархічної кластеризації

Ключові слова: інтелектуальний аналіз текстів, аналіз формальних понять, кластеризація, семантичні поля

В работе предложена модель семантического контекста текстовых массивов. Показано, что на основании решетки семантических концептов можно сформировать семантический базис для группирования текстов с помощью иерархической кластеризации

Ключевые слова: интеллектуальный анализ текстов, анализ формальных понятий, кластеризация, семантические поля

The semantic context model of text arrays has been suggested in this work. It is shown that the semantic basis for texts grouping using hierarchical clusterization can be formed on the base of semantic concepts lattice

Key words: text mining, formal concepts analysis, clusterization, semantic fields

УДК 519.765:519.767:004.93

ГРУПУВАННЯ ТЕКСТОВИХ ДАНИХ НА ОСНОВІ МОДЕЛІ СЕМАНТИЧНОГО КОНТЕКСТУ

Б.М. Павлишенко
Кандидат фізико-математичних наук, доцент
Факультет електроніки
Львівський національний університет ім. І. Франка.
Львів, вул. Драгоманова, 50, 79005
Контактний тел.: 0505037290
e-mail: pavlsh@yahoo.com

Постановка проблеми

Одним із важливих напрямів сучасних інформаційних технологій є інтелектуальний аналіз текстових даних [1,2]. В такому аналізі часто використовують кластерний аналіз, за допомогою якого групують текстові документи із спільними характеристиками. Кластеризація текстових документів відбувається у багатомірному векторному просторі, кожний вимір якого відповідає квантитативній характеристиці лексеми зі словників аналізованих текстових масивів [1,2,3]. Такою характеристикою може бути, наприклад, текстова частота лексеми. Ефективним методом аналізу даних є також теорія аналізу формальних концептів [4,5,6]. В цій теорії розглядається відношення об'єктів та їх атрибутів, на основі якого будують алгебраїчну решітку формальних концептів. Кожний концепт об'єднує множину об'єктів та їх спільних атрибутів. На основі частих множин спільних атрибутів виявляють асоціативні правила, які відображають зв'язки між атрибутами на множині аналізованих об'єктів. Перспективним для аналізу текстових даних є об'єднання методів кластеризації та аналізу формальних понять. Зокрема, методи аналізу формальних понять можуть бути використані для формування семантичного базису векторного простору, в якому кластеризуються текстові документи.

Аналіз останніх досліджень та публікацій

Кластерний аналіз є ефективним при вивченні структури текстових масивів [1,2,3]. Для представлення текстових документів часто використовують модель векторного простору [2]. Текстовий масив можна представити у вигляді матриці слів та документів, в якій колонки визначають документи, а рядки – частоти лексем в цих документах. Тоді кожна колонка є вектором частот лексем для заданого документа, який задається номером колонки. Мірою відстані між двома документами може бути кут між векторами цих документів в утвореному векторному просторі. Такий підхід має також ряд проблем, зокрема, розмірність аналізованого простору є великою, оскільки зумовлена розміром словника. Документи також можуть бути квантитативно близькими не тільки за частотами окремих лексем, а також за характеристиками заданих лексемних об'єднань, наприклад, семантичних полів [7,8]. Пошук комплексних характеристик текстових документів є важливим, зокрема при аналізі авторства текстів, так як лексемний частотний спектр творів може бути однаковим, але відрізнятись за характеристиками комбінованих лексемних груп. В теорії аналізу формальних понять (Fomal Concept Analysis) [4,5,6] аналізують ієрархії формальних понять використовуючи математичний апарат теорії алгебраїчних решіток. Однією із актуальних проблем є побудова моделі формального контексту для семантичних характеристик текстових даних на основі векторної моделі текстових документів та формального аналізу понять.

Цілі статті

Для виявлення нових підмножин метаданих, які будуть ефективними в алгоритмах аналізу текстових

масивів, розглянемо структурний поділ лексемного складу за семантичними полями. Для аналізу семантичного простору побудуємо теоретико-множинну модель семантичних полів. Розглянемо модель формального семантичного контексту текстових масивів. Проаналізуємо алгебрічну решітку семантичних концептів. На основі змісту концептів, які відображають тематику аналізу побудуємо тематичне семантичне поле. Лексемний склад цього поля використаємо як базис векторного простору, в якому можна реалізувати кластеризацію текстових документів.

Основний матеріал. Теоретико-множинна модель текстових даних.

Розглянемо модель, яка описує сукупність текстових документів, лексемний склад та семантичні поля. Нехай існує деякий словник лексем, які зустрічаються у текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \left\{ w_i \,|\, i = 1,2...,N_w \right\} \tag{1}$$

Сукупність текстових документів опишемо такою множиною

$$D = \left\{ d_j \,|\, j = 1,2...,N_d \right\} \tag{2}$$

Введемо множину семантичних полів

$$S = \left\{ s_k \,|\, k = 1,2...,N_s \right\} \tag{3}$$

Введемо відображення лексемного складу словника W на множину семантичних полів S за допомогою деякого оператора Uws

$$U_{ws} : w_i \to s_k,\ i = 1,2...,N_w; k = 1,2...,N_s \tag{4}$$

Оператор Uws задамо таблицею, яка визначається експертним лексикографічним аналізом [7,8]. Лексемний склад семантичного поля sk визначимо як

$$W_k^s = \left\{ w_i \,|\, w_i \overset{U_{ws}}{\to} s_k, i = 1,2...,N_w \right\} \tag{5}$$

Введемо матрицю семантичних ознак типу "частоти_семантичних_полів–документи"

$$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1,j=1}^{N_s,N_d} \tag{6}$$

де p_{kj}^{sd} - частота семантичного поля Sk в лексемному складі документа dj , яку обчислимо за формулою

$$p_{kj}^{sd} = \frac{n_{kj}^{sd}}{N_j^t} \tag{7}$$

де n_{kj}^{sd} - кількість лексем семантичного поля Sk в лексемному складі документа dj . Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right) \tag{8}$$

відображає документ dj в Ns-мірному семантичному просторі текстових документів. Розглянемо бінарні семантичні характеристики текстового документа

$$p_{kj}^{bs} = \begin{cases} 1, & p_{kj}^{sd} \ge \overline{p}_k^{sd} \\ 0, & p_{kj}^{sd} < \overline{p}_k^{sd} \end{cases}, \tag{9}$$

де $\overline{p}_k^{sd}$ - деяке порогове значення частоти семантичного поля Sk. Враховуючи (9), вектор бінарних семантичних характеристик можна записати у вигляді

$$V_j^{bs} = \left(p_{1j}^{bs}, p_{2j}^{bs}, ..., p_{N_s j}^{bs} \right). \tag{10}$$

Модель решітки семантичних концептів

Розглянемо модель семантичної структури текстових масивів використовуючи теорію аналізу формальних понять [4,5,6]. Визначимо семантичний контекст як трійку

$$K_s = (D,S,I), \tag{11}$$

де D – масив документів, S - множина семантичних полів, I - відношення належності семантичного поля до даного документу

$$I \subseteq D \times S,\ I = \left\{ (d_i, s_k) \right\}. \tag{12}$$

Пара (di, Sk) означає, що документ di характеризується семантичним полем Sk, тобто $p_{kj}^{bs} = 1$.

Уведемо решітку семантичних концептів. Для деяких $Ext \subseteq D,\ Int \subseteq S$ визначимо такі відображення

$$\begin{aligned} Ext' &= \left\{ s \in S \,|\, d \in Ext : dIs \right\} \\ Int' &= \left\{ d \in D \,|\, s \in Int : dIs \right\} \end{aligned} \tag{13}$$

Множина Ext′ описує семантичні поля, які властиві документам множини Ext, а множина Int′ описує документи, які володіють семантичними полями множини Int. Уведемо семантичний концепт як пару

$$Concept = (Ext, Int) \tag{14}$$

до якої належать лексеми з множини $Ext \subseteq D$ та семантичні поля з множини $Int \subseteq S$ з такими умовами

$$\begin{cases} Ext' = Int, \\ Int' = Ext. \end{cases} \tag{15}$$

Множину Ext назвемо об'ємом, а Int– змістом семантичного концепту Concept. В семантичному контексті утворюється частково-впорядкована множина семантичних концептів

$$\psi(D,S,I) = \{ Concept_m = (Ext_m, Int_m) \,|\, m = 1,2,...N_{ct} \} \tag{16}$$

де Nct – кількість виявлених семантичних концептів у формальному семантичному контексті масиву текстових документів. Семантичний концепт

$$\text{Concept}_1 = (\text{Ext}_1, \text{Int}_1) \tag{17}$$

є менш загальним за об'ємом чим концепт

$$\text{Concept}_2 = (\text{Ext}_2, \text{Int}_2) \tag{18}$$

тобто виконується умова

$$(\text{Ext}_1, \text{Int}_1) \le (\text{Ext}_2, \text{Int}_2), \tag{19}$$

якщо

$$\text{Ext}_1 \subseteq \text{Ext}_2 \;\Leftrightarrow\; \text{Int}_1 \supseteq \text{Int}_2. \tag{20}$$

В цьому випадку концепт Concept2 можна вважати узагальненням концепту Concept1. Семантичний концепт можна розглядати як підматрицю семантичного контексту, яка повністю заповнена одиницями. Решітку концептів часто відображають за допомогою діаграм Гассе. В аналізі семантичного контексту кожний елемент діаграми представляє семантичний концепт. На верхньому рівні діаграми концепт включає в себе всі текстові документи і нульову множину семантичних полів. На другому рівні в елементи діаграми входить одне семантичне поле, на третьому – два семантичних поля і так до найнижчого рівня, який включає в себе всі семантичні поля та нульову множину текстових документів. Такі діаграми відображають внутрішню семантичну структурну організацію масивів текстових документів на основі теорії формального аналізу понять.

Кластеризація текстових документів

Семантичні концепти об'єднують групи текстових документів та семантичні поля, які є властиві цим документам. У випадку тематичного аналізу текстових даних в решітці семантичних концептів можна виявити підмножину змістів концептів {Intj}, які будуть відображати тематику аналізу. Тематичне семантичне поле розглянемо як об'єднання змістів таких концептів:

$$S_t = \{\, s_i \mid s_i \in \text{Int}_j \,\} \tag{21}$$

Розглянемо векторний простір, базис якого утворюється на основі елементів множини St. В такому просторі кожен текстовий документ буде розглядатись як вектор

$$V_j^{td} = \left(p_{1j}^{ts}, p_{2j}^{ts}, \ldots, p_{N_{St}j}^{ts}\right) \tag{22}$$

де NSt- кількість семантичних полів у тематичному полі St. На відміну від векторного представлення документів (10), у цьому векторному представленні присутні частоти лише деякої підмножини семантичних полів, які відображають тематику заданого аналізу. Векторне представлення документа (22) використаємо для групування документів за допомогою ієрархічної кластеризації, яка дасть можливість виявити групи документів, які є близькими за визначеною тематикою. Такий підхід є ефективніший ніж кластеризація за наперед визначеною множиною семантичних полів, оскільки тематично близькі документи можуть сильно відрізнятись за несуттєвими семантичними полями і отже не попадуть в спільний кластер.

Розглянемо групування документів за семантичними ознаками за допомогою алгоритму ієрархічної кластеризації. Нехай є множина текстових документів D, яка описується виразом (2) та множина кластерів

$$C = \{\, c_m \mid m = 0,1,2\ldots, N_c \,\} \tag{23}$$

Необхідно побудувати відображення множини документів на множину кластерів :

$$U_{DC} : D \to C \tag{24}$$

Відображення UDC задає модель даних, яка є розв'язком задачі кластеризації [1,2,3]. Кожний елемент cm множини кластерів C складається з підмножини текстових документів, які подібні між собою відповідно до деякої кількісної міри подібності r

$$c_m = \{\, d_i, d_j \mid d_i \in D,\ d_j \in D,\ r(d_i, d_j) < \varepsilon \,\}, \tag{25}$$

де ε - визначає деякий поріг для включення документів в кластер. Величина r(di, dj) є відстанню між елементами di та dj . Якщо виконується умова

$$r(d_i, d_j) < \varepsilon \tag{26}$$

то елементи вибірки вважають подібними і приналежними до спільного кластера. В іншому випадку елементи знаходяться у різних кластерах. Матриця

$$R = \{\, r_{ij} = r(d_i, d_j) \,\} \tag{27}$$

є матрицею відмінностей в алгоритмі кластеризації. Очевидно, що діагональні елементи цієї матриці дорівнюють нулю. Оскільки на множині текстових документів введено поняття відстані, то кожен документ представляють у вигляді точки в Ns–мірному просторі R^{Ns} семантичних полів. Є декілька методів обрахунку мір близькості точок в Ns–мірному просторі, зокрема, евклідова відстань обраховується так

$$r_e(d_i, d_j) = \sqrt{\sum_{k=1}^{N_s} \left(p_{ki}^{sd} - p_{kj}^{sd}\right)^2}. \tag{28}$$

Подібність між двома текстовими документами в Ns–мірному просторі також визначається кутом між векторами цих документів і за кількісну міру можна взяти косинус цього кута.

Розглянемо ієрархічний метод агломеративної кластеризації. На першому кроці вся множина текстових документів розглядається як множина кластерів:

$$c_1 = \{\, d_1 \,\}, c_1 = \{\, d_1 \,\}, \ldots c_{Nd} = \{\, d_{Nd} \,\}, \tag{29}$$

На наступному кроці два близьких один до одного документа (наприклад dp і dq) об'єднуються в один

спільний кластер, нова множина на цьому кроці вже складається із Nd-1 кластерів і має вигляд

$$c_1 = \{ d_1 \}, c_2 = \{ d_2 \}, ... c_p = \{ d_p, d_q \} ... c_{Nd-1} = \{ d_{Nd-1} \}, \quad (30)$$

Повторюючи кроки, на яких будуть об'єднуватися кластери, отримаємо множину із Nc кластерів. Процес об'єднання кластерів завершується на тому кроці алгоритму, коли жодна пара кластерів не відповідає порогу об'єднання для міри близькості елементів. На кожній ітерації алгоритму необхідно робити перерахунок між кластерами. Враховуючи те, що кластери можуть складатися з декількох об'єктів, існують різні методи формування та об'єднання кластерів на основі відстаней між об'єктами всередині кластера. Наприклад, метод найближчого сусіда полягає у виборі найменшої відстані між двома кластерами p і q :

$$r(p,q) = \min \{ r(d_{pi}, d_{qj}), i \in (1,2,...,N_p), j \in (1,2,...,N_q) \} \quad (31)$$

Використовуючи наведені кроки ієрархічної кластеризації отримаємо кластерну структуру текстових документів в просторі семантичних полів. Базис цього простору буде визначатись змістом семантичних концептів текстових документів.

Висновки

Запропонована в роботі модель семантичного контексту відображає структурну семантичну організацію текстових масивів. В семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об'єм – масивами текстових документів. Побудова решітки семантичних концептів в текстових документах дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи текстових документів, які об'єднані спільною групою семантичних ознак. На основі змістів концептів, які відповідають заданій тематиці можна сформувати базис семантичного простору текстових документів. Ієрархічна кластеризація документів у такому просторі дає можливість згрупувати у спільних кластерах тематично близькі документи та ігнорувати відмінності за несуттєвими для тематики семантичними полями.

Література

1. Брасегян А.А. Анализ данных и процессов: учеб. Пособие / А.А.Брасегян, М.С.Куприянов, И.И.Холод, М.Д.Тесс, С.И.Елизаров.–СПб.:БХВ–Петербург,2009.-512с.:ил.
2. Pantel P. From Frequency to Meaning: Vector Space Models of Semantics / Patrick Pantel, Peter D. Turney // Journal of Artificial Intelligence Research.–2010.–vol.37.–pp.141-188.
3. Жамбю М. Иерархический кластер-анализ и соответсвия: пер. с фр. – М.: Финансы и статистика, 1988. – 342 с: ил.
4. Ganter B. Formal Concept Analysis: Mathematical Foundations/ B.Ganter, R.Wille.–Springer, 1999.
5. Kuznetsov S.O. Comparing Performance of Algorithms for Generating Concept Lattices / S.O. Kuznetsov, S.A. Obiedkov //Journal of Experimental and Theoretical Artificial Intelligence.–2002.–vol.14.–pp.189-216.
6. Cimiano P. Learning Concept Hierarchies from Text Corpora, using Formal Concept Analysis / P. Cimiano, A. Hotho, S. Staab // Journal of Artificial Intelligence Research.–2005.–vol.24.–pp.305-339.
7. Вердиева З.Н. Семантические поля в соверменном английском языке/ З.Н. Вердиева –М.: Высшая школа,1986.–120с.
8. Левицкий В.В. Экспериментальные методы в семасиологии/ В.В. Левицкий, И.А. Стернин. – Воронеж: Изд-во ВГУ, 1989. – 192с.